*W336–W339* *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
*DOI: 10.1093/nar/gkh365*

# Phydbac2: improved inference of gene function using interactive phylogenomic profiling and chromosomal location analysis

**François Enault*, Karsten Suhre, Olivier Poirot, Chantal Abergel and Jean-Michel Claverie**

Information Génomique & Structurale (UPR CNRS 2589), Institut de Biologie Structurale et Microbiologie,
31, chemin Joseph Aiguier, 13402 Marseille Cedex 20, France

Received February 14, 2004; Accepted March 4, 2004

## ABSTRACT

**Phydbac (phylogenomic display of bacterial genes) implemented a method of phylogenomic profiling using a distance measure based on normalized BLAST scores. This method was able to increase the predictive power of phylogenomic profiling by about 25% when compared to the classical approach based on Hamming distances. Here we present a major extension of Phydbac (named here Phydbac2), that extends both the concept and the functionality of the original web-service. While phylogenomic profiles remain the central focus of Phydbac2, it now integrates chromosomal proximity and gene fusion analyses as two additional non-similarity-based indicators for inferring pairwise gene functional relationships. Moreover, all presently available (January 2004) fully sequenced bacterial genomes and those of three lower eukaryotes are now included in the profiling process, thus increasing the initial number of reference genomes (71 in Phydbac) to 150 in Phydbac2. Using the KEGG metabolic pathway database as a benchmark, we show that the predictive power of Phydbac2 is improved by 27% over the previous version. This gain is accounted for on one hand, by the increased number of reference genomes (11%) and on the other hand, as a result of including chromosomal proximity into the distance measure (16%). The expanded functionality of Phydbac2 now allows the user to query more than 50 different genomes, including at least one member of each major bacterial group, most major pathogens and potential bio-terrorism agents. The search for co-evolving genes based on consensus profiles from multiple organisms, the display of Phydbac2 profiles side by side with COG information, the inclusion of KEGG metabolic pathway maps the production of chromosomal proximity maps, and the possibility of collecting and processing results from different Phydbac queries in a common shopping cart are the main new features of Phydbac2. The Phydbac2 web server is available at http://igs-server.cnrs-mrs.fr/phydbac/.**

## INTRODUCTION

Three major non-similarity-based methods for the prediction of bacterial pairwise gene relationships have been proposed in the past (1–6) and are now widely established. These methods are based on (i) the co-evolution of functionally related genes (phylogenomic profiles), (ii) the conservation of chromosomal proximity (operon-like structures) and (iii) the detection of gene fusion events (Rosetta-stones). A number of web-servers are presently available that implement these methods at different levels of detail, such as STRING (7). The primary data on gene-absence/presence used by STRING is derived from the well-established ‘Cluster of Orthologous Groups’ (COG) database (8). We previously introduced the web server Phydbac (phylogenomic display of bacterial genes) (9) implementing a new method of phylogenomic profiling based on normalized BLAST (10) scores rather than on the simple presence or absence of a given gene in a genome (as for instance in STRING). This approach was shown to improve the predictive power of phylogenomic profiling by about 25% with respect to the classical approach (11). However, the original Phydbac was limited to queries for genes from *Escherichia coli* and its predictions solely based on phylogenomic profiling. Here we describe Phydbac2, which removes these limitations by allowing 50 genomes to be queried and integrates chromosomal proximity and gene-fusion analysis to generate improved functional predictions. Numerous additional new features (described below) were also implemented to make Phydbac2 an even more useful tool for the inference of bacterial gene function.

*To whom correspondence should be addressed. Tel: +33491164548; Fax: +33491164549; Email: enault@igs.cnrs-mrs.fr

## NEW FEATURES IN PHYDBAC2

The concept of *Consensus Phylogenomic Profiles* (*CPP*) is a major addition to the original Phydbac. The CPP is automatically computed from the displayed individual gene profiles by simply averaging the scores in each column where at least half of the selected genes show a significant similarity (Figure 1). These profiles can be generated from a query gene and its co-evolving neighbors, or from the genes involved in a predefined KEGG pathway (12). The CPP can then be used as a query, in which all selected genes contribute equally. Such a search will identify target genes exhibiting a co-evolution relationship to the whole pathway, rather than to a specific gene.

A different type of CPP is also generated by clicking on the individual phylogenomic profile of any given gene. This will pop out the profiles (and the CPP) for the corresponding genes (putative orthologs) in all other bacterial genomes. This CPP now has the advantage of being independent of the starting genome. Such an ortholog-average CPP usually amplifies the noisy phylogenomic signal obtained when the similarity between the query gene and its best match in other genomes is low. It often leads to the identification of more functionally coherent neighborhoods than profiles computed on individual genes.

*The analysis of conserved gene to gene chromosomal proximity* is now implemented. Co-localized genes were initially detected by distributing the 150 bacterial genomes (including multiple strains of the same bacteria or evolutionary close species) into 87 groups (clusters of very similar organisms) on the basis of the multiple alignments of the 150 homologs of three genes. Two genes were then classified as 'co-localized' when their most similar sequences were found at a distance <2000 bp in three (or more) genomes of the 87 groups. For all the genes satisfying this constraint, the chromosome regions of the relevant organisms can be displayed, with each of the regions anchored on the position of the gene for which the map was requested. The positions of the genes found to co-localize with the query are highlighted in different colors (Figure 1).

*Gene fusion* events are derived from the FusionDB database (13) and are flagged in the Phydbac2 output for all genes in a phylogenomic neighborhood that exhibits Rosetta-stone sequences in one or several genomes. The Phydbac2 output now provides active links to the detailed analyses of these fusion events, including multiple alignments and phylogenetic trees. The 'reality' of a putative gene fusion event can thus be checked interactively.

*The improved predictive power* of Phydbac2 with respect to the original Phydbac is due to two factors: (i) the higher

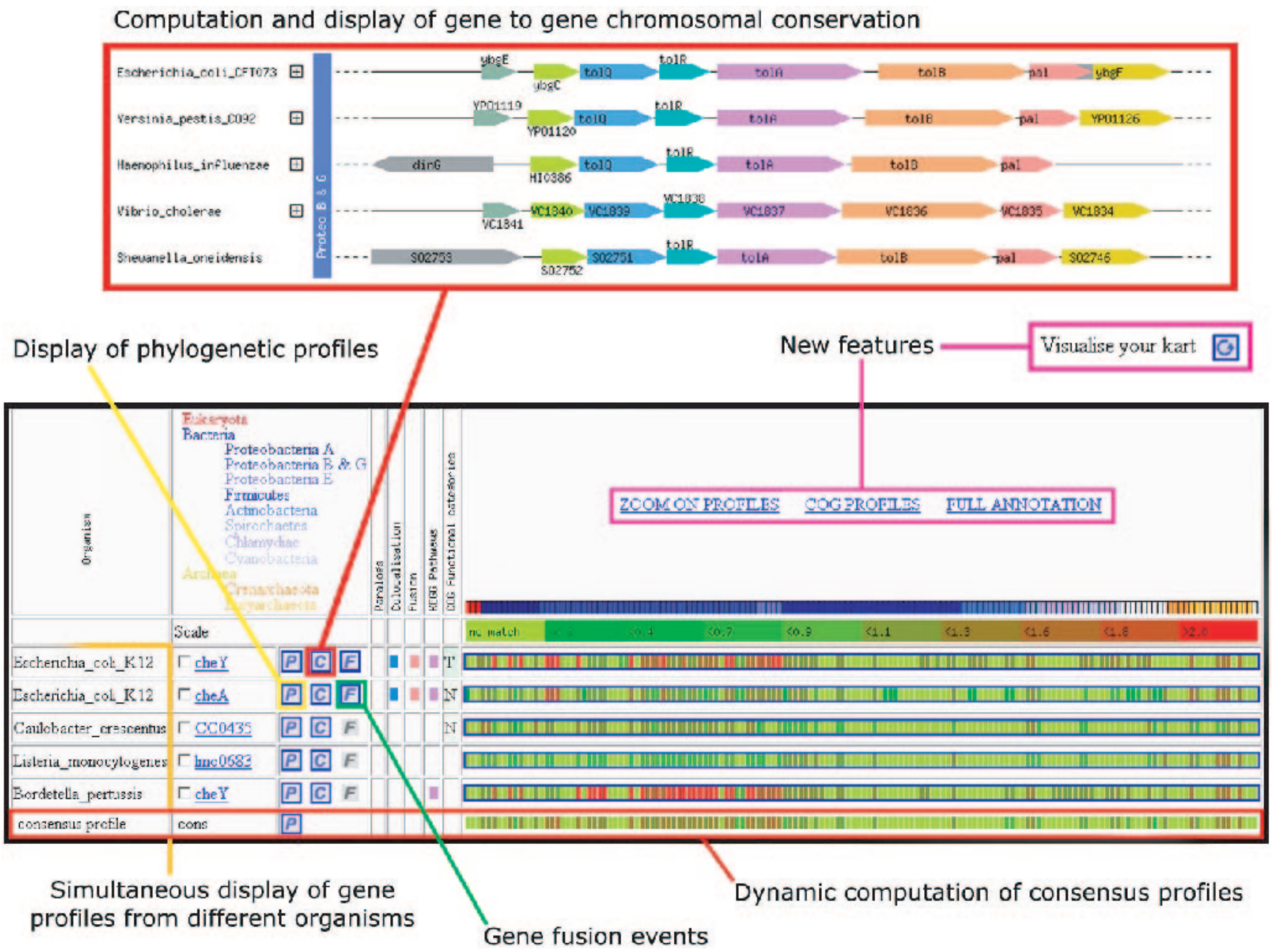

**Figure 1.** Screenshots of Phydbac2 typical outputs and new features.

number of profiled genomes (150 instead of 71), and (ii) the integration of the chromosomal proximity in the computation of the pairwise gene distance used for the functional predictions. We quantified this improvement using the KEGG metabolic pathway database (13) as reference. Assuming that genes belonging to the same KEGG pathway should have a higher probability of co-evolution than unrelated ones, we expect improved phylogenomic profiles to identify a higher number of such gene pairs in their respective neighborhoods. For this validation, the KEGG pathways of 11 organisms (chosen to represent all main bacterial clades) were used, including 870 sets of pathways, for a total of 11 643 gene occurrences. For each gene in that data set, and using a fixed number of nearest phylogenomic neighbors, we computed the fraction of genes belonging to the query neighborhood *and* sharing the query KEGG pathway. The results are shown in Figure 2 as a function of the neighborhood size. When considering the ten nearest neighbors, Phydbac2 outperforms its predecessor by about 11%. Integrating chromosomal proximity to the distance measure yields a supplementary 16% gain in performance. Overall, Phydbac2 exhibits a 27% improvement in predictive power over the original Phydbac.

*Automatic gene annotation* using the KEGG database is available through the 'Full annotation' link. As described in detail in (11), for a given gene and for different neighborhood sizes we determined the number of genes sharing the same KEGG pathway. We then computed the probability of observing as many or more genes from the same pathways through random samples of the same size. In cases associated with probability less than 1/100 000, the given gene is deemed to have a co-evolutionary (and most likely functional) relationship with the relevant KEGG pathway and the association is reported together with the corresponding statistics (probability and neighborhood size). As a test, we automatically annotated all genes occurring in the KEGG pathways to see how much of the information would be recovered. Using the above probability threshold (1E-5) and considering neighborhoods of up to 500 genes we could retrieve 1093 of the 2149 KEGG annotations of *Escherichia coli* (793 correct ones when limiting the neighborhood to 10 neighbors and 383 when using a threshold of 1E-10).

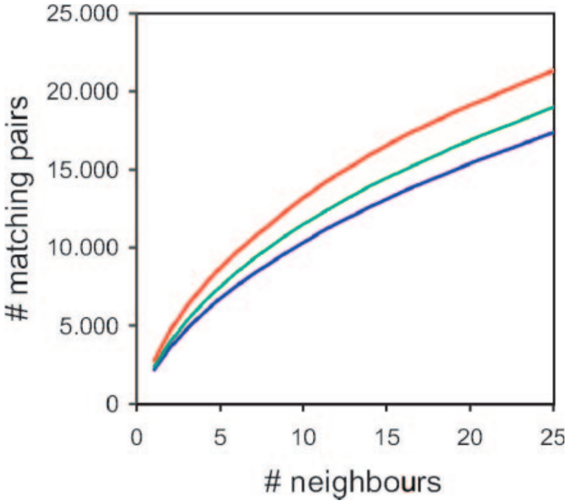

**Figure 2.** Number of predicted matching pairs for a given number of closest neighbors: old Phydbac (blue), new Phydbac2 (green), new Phydbac2 + chromosomal proximity (red).

Finally, Phydbac2 allows the *display of COG annotation* on phylogenomic profiles. This new function is useful to highlight potential inconsistencies between the COG annotation of a query gene and that of its best match in a target genome. Such a display also provides a visual illustration of the matches taken into account by Phydbac that are not (or are mis-) assigned to a COG (i.e. multiple domain protein, or highly divergent homologs).

## CONCLUDING REMARKS

Due to size limitation, this article could provide only a brief overview of the new functions offered by Phydbac2. They are described in much greater details in an animated guided tour (best viewed with the Flash media player plugin) available on the Phydbac2 website. A static HTML version is also provided. Among the most useful new features there is now the possibility of collecting genes from different queries in a shopping cart (e.g. for consensus profile searching and phylogenomic-distance-based tree drawing), the display of additional information such as KEGG pathway maps and other detailed annotations (paralogs, fusion events, co-localization, etc.). The processing of entire genomes for their inclusion in the web server being now completely automated, additional bacteria will be added continuously as their genome sequences become publicly available. Specific strains or species that are not yet included on the web server may also be added upon email request.

## ACKNOWLEDGEMENTS

KEGG metabolic pathway maps and data were downloaded from http://www.genome.ad.jp/kegg/.COG annotations were obtained from RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/).

## REFERENCES

1. Galperin,M.Y. and Koonin,E.V. (2000) Who's your neighbor? New computational approaches for functional genomics. *Nat. Biotechnol.*, **18**, 609–613.
2. Sali,A. (1999) Functional links between proteins. *Nature*, **402**, 23–26.
3. Marcotte,E.M. (2000) Computational genetics: finding protein function by nonhomology methods. *Curr. Opin. Struct. Biol.*, **10**, 359–365.
4. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci.*, **96**, 4285–4288.
5. Dandekar,T., Snel,B., Huynen,M. and Bork,P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, **23**, 324–328.
6. Enright,A.J. and Ouzounis,C.A. (2001) Functional associations of proteins in entire genomes via exhaustive detection of gene fusion. *Genome Biol.*, **2**, 341–347.
7. von Mering,C., Huynen,M., Jaeggi,D., Schmidt,S., Bork,P. and Snel,B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.*, **31**, 258–261.
8. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new

developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.

9. Enault,F., Suhre,K., Poirot,O., Abergel,C. and Claverie,J.M. (2003) Phydbac (phylogenomic display of bacterial genes): an interactive resource for the annotation of bacterial genomes. *Nucleic Acids Res.*, **231**, 3720–3722.
10. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
11. Enault,F., Suhre,K., Poirot,O., Abergel,C. and Claverie,J.M. (2003) Annotation of bacterial genomes using improved phylogenomic profiles. *Bioinformatics*, **19**, i105–i107.
12. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
13. Suhre,K. and Claverie,J.M. (2004) FusionDB: a database for in-depth analysis of prokaryotic gene fusion events. *Nucleic Acids Res.*, **32**, D273–D276.